Research Memorandum 78-12

SCORE QUALITY ISSUES RELATED TO INDIVIDUAL AND WEAPON CREW CRITERION-REFERENCED PERFORMANCE TESTS

Frederick Steinheiser, Jr.
Army Research Institute for the Behavioral and Social Sciences

and

C. Wesley Snyder, Jr.
American Institutes for Research

UNIT TRAINING AND EVALUATION SYSTEMS TECHNICAL AREA

U. S. Army
Research Institute for the Behavioral and Social Sciences

April 1978

Army Project Number
(16) 2Q762722A764

Unit Training Standards
& Evaluation

(9) Research ~~Memorandum 78-12~~ memo, (12) 19

(6) SCORE QUALITY ISSUES RELATED TO INDIVIDUAL AND WEAPON CREW CRITERION-REFERENCED PERFORMANCE TESTS.

(10) Frederick/Steinheiser, Jr. C. Wesley/Snyder, Jr
Army Research Institute for the Behavioral and Social Sciences

and

C. Wesley Snyder, Jr.
American Institutes for Research

(14) ARI-RM-78-12

Angelo Mirabella, Work Unit Leader

Submitted by:
Frank J. Harris, Chief
UNIT TRAINING AND EVALUATION SYSTEMS TECHNICAL AREA

(11) Apr [redacted]78

Approved by:

A. H. Birnbaum, Acting Director
Organizations and Systems
Research Laboratory

Joseph Zeidner, Acting Technical
Director
U.S. Army Research Institute for
the Behavioral and Social Sciences

Research Memorandums are informal reports on technical research problems. Limited distribution is made, primarily to personnel engaged in research for the Army Research Institute.

Presented at the Military Testing Association Conference, October 1976.*

Score Quality Issues Related to Individual and Weapon Crew Criterion-Referenced Performance Tests

Frederick Steinheiser, Jr.
Army Research Institute for the
Behavioral and Social Sciences
Alexandria, Virginia

and

C. Wesley Snyder, Jr.
American Institutes for Research
Washington, D.C.

In criterion-referenced personnel testing, an examinee is advanced if he is able to respond correctly to at least some specified percentage of the items or trials comprising the test domain. This minimum passing percentage, or "criterion level," reflects the degree of mastery deemed sufficient to pass a specific instructional or training objective. The actual percentage of items that he would correctly respond to in the population of items may be termed his "true level of functioning" (or "true state of combat readiness"). In practice, the advancement-retention decision must be made from a fairly small sample of items. As a result, errors in the classification of examinees are unavoidable. The task of the examiner is to minimize these errors as much as possible through good test design.

Due to a variety of practical constraints, such as time, manpower, and cost of testing, test length is an important consideration. A reasonable goal for the decisionmaker is to be able to classify examinees into groups of masters and nonmasters (or combat ready vs. not combat ready) as accurately as possible, without requiring excessive numbers of test items or trials. This problem is addressed by the first two models presented in this paper: (i) the binomial model, which provides the probability of an

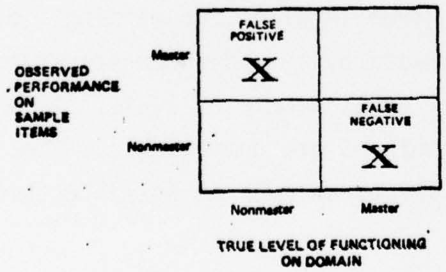
*The views expressed in this paper are those of the authors and do not necessarily imply endorsement by the U.S. Army.

examinee obtaining a test score given his assumed true level of functioning; and (ii) the Bayesian model, which provides a probabilistic estimate of an examinee's true level of functioning given his test score.

Another practical constraint in personnel evaluation concerns the heterogeneity of items on different tests, and the heterogeneity of various examinee populations. Especially in military testing, it may not be possible to give each examinee group the same set of items or trials. The third section of this paper presents a model which allows examinees of differing skill levels to be scaled on a common skill metric, even when they have not all taken the same set of test items.

Binomial Model

A marksman or tank crew unit is assumed to have a true level of functioning (i.e., when tested on the entire domain, the examinee would "pass" a certain proportion of items). Based on a hypothetical standard involving some proportion of items passed, we can split the examinees into two groups--masters and nonmasters. When actually tested on a subset of the domain (e.g., 18 items) with a specified mastery level (e.g., 83 percent), we can identify masters (e.g., those who pass 15 out of 18 items) and nonmasters. The situation is diagrammed below.


FALSE POSITIVE
X
Master
OBSERVED PERFORMANCE ON SAMPLE ITEMS
FALSE NEGATIVE
X
Nonmaster
Nonmaster
Master
TRUE LEVEL OF FUNCTIONING ON DOMAIN

In each of the four boxes an inference can be drawn about the true level of functioning given performance on a sample of items or about performance, given true skill. The two boxes marked with an "X" indicate errors on the part of the examiner. First, he might pass examinees whose true level of functioning is below the specified mastery level (called a false positive error). Alternatively, he might fail examinees whose true level of functioning is at or above the specified mastery level (called a false negative error). The probability of these errors occurring is a function of the number of items included in the test. The binomial model serves as an analytic aid to the consideration of the test length issue in terms of probable misclassification, provided the decisionmaker is willing to specify true skill levels for the intended examinee population.

In order to calculate the expected error in decisionmaking, we need to specify the lowest proficiency level required for an examinee to be considered a master (called minimal proficiency level) and the highest proficiency level that an examinee could attain and still be considered a nonmaster (called maximum nonproficient level). These levels would intuitively be at the same point except that the test is not completely accurate. Given these parameters as set by the decisionmaker based on his knowledge of the situation, the probability of false positive and false negative errors for minimal masters and maximal nonmasters can be calculated for any given test length and cutting score.

For example, assume that a true proficiency (which cannot be directly observed) of 90 percent is set as the minimal proficiency level, and that a true proficiency of 70 percent is set as the maximum nonproficient (not ready for combat) level. The region between these two cutoff scores is an area of uncertainty. Perhaps if an examinee's score fell in this region, he would receive a brief course for refresher training, whereas if his score were 70 percent, he would receive extensive retraining.

Continuing with this example, we also need to specify values for classification errors. Suppose that the examiner is unwilling to accept more than 25 percent of the examinees whose true proficiency state is 70 percent; and he wants to reject no more than 25 percent of those whose true proficiency is at 90 percent. Thus, the probabilities for false positive and false negative misclassifications have been set before any data have been collected, at .25 each. Given the values of lower bound for true proficiency (90 percent), upper bound for true non-proficiency (70 percent), and tolerable false positive and false negative error rates (.25 each), it is then possible to determine the minimum number of trials that should be given and how many trials should be correct. Table 1 shows, for example, that for an observed score of 80 percent, at least nine trials must be given and at least eight "hits" must be made in order to evaluate a crew as combat ready, subject to the above constraints. (The computational details may be found in an Army Research Institute paper by Macready, Epstein, Steinheiser, and Mirabella, 1978.)

Table 1. Probability of misclassification* as a function of the length of Table VIII.

Required Mastery Level	Table VIII Length	Passing Score	True Level of Tank Crew Functioning				
			50%	60%	70%	80%	90%
80%	36	29	–	1	11	43	2
	18	15	–	3	16	50	10
	9	8	2	7	20	56	23
70%	36	26	1	9	53	9	–
	18	13	5	21	47	13	1
	9	7	9	23	54	26	5
60%	36	22	12	48	9	–	–
	18	11	24	44	14	2	–
	9	6	25	52	27	9	1

* Entries to the left of dotted lines are probabilities (percent) of false positives; those to the right of dotted lines are false negatives.

(Adapted from Millman, 1973 and Novick and Lewis, 1974)

Another way of using the binomial model is to consider error rates for different cutoff points and lengths of an existing test. The present Army Table VIII for Tank Gunnery Qualification consists of 18 items. Inspection of Table 1 indicates that for an 80 percent cutoff level, this test length decreases the misclassifications for the 70 percent and 90 percent level to .16 and .10, respectively. Doubling the test to 36 items further reduces the misclassification to .11 and .02, respectively. The decisionmaker can assess the trade-offs of test-length and misclassifications in terms of the practical constraints of his test situation in order to arrive at an optimal test design.

It is obvious that in any evaluation program, some people who are qualified will be classified as not (yet) qualified, and vice versa. There is no such thing as giving an examinee population a set of 10, 100, or even 1,000 trials and hoping to be able to exactly separate the "sheep from the goats." But this binomial model quantifies the worst level of misclassifications that might be expected for any observed proficiency level, when test length and cutoff scores have been specified by the examiner.

Bayesian Model

Whereas the binomial model provides us with information about potentially observable behavior based on the decisionmaker's best guesses about the true skill levels in the intended examinee population, the Bayesian model improves the classification of observed performance. Turning it to the advantage of the test developer, the Bayesian model can be used to evaluate the usefulness of what is already known about the examinee population to improve the situation outlined by the binomial model.

Consider a testing program which classifies examinees into three combat readiness (proficiency) states, quantifiable as percentages: 90 percent indicates combat ready, 80 percent indicates a need for brief refresher training, and 70 percent indicates a need for extensive retraining. Suppose that the probability of getting a tank crew with a 90 percent readiness state is thought to be .60. This means that on the basis of historical information, 60 percent of all tested examinees has been found to have a readiness state of 90 percent. Also suppose that the probability of observing an 80 percent proficient crew is 30 percent, and the probability of observing a 70 percent proficient crew is 10 percent. (The examiner's prior information about the proficiency distribution of the examinee population is represented by .6, .3, and .1.)

If we observe that an examinee scores 13 hits out of 18 trials, what is the probability that this examinee is combat ready, or needs brief refresher training, or needs extensive retraining? Note that on the basis of the observed data, 13/18 = 72 percent, which would seem to indicate a high probability for need of extensive retraining.

The test data can be combined with the prior information in the following way. The sampling distribution of the number of hits in 18 trials, for any proficiency, may conveniently be assumed to be a binomial distribution. It is then possible to go to a table of binomial probabilities to find the likelihood of getting 13 hits in 18 attempts with a 90 percent chance of success; 13 out of 18 with an 80 percent chance of success; and 13 out of 18 with a 70 percent chance for success. The values for these likelihoods are shown in column 3 of Table 2.

Next, we multiply the prior probability (column 2) times the likelihood value for each row. These values are shown in column 4. The sum of these values in column 4 equals .0785. The column 5 (posterior) probabilities are obtained by dividing each column 4 row entry by the column 4 total. The entries in column 5 represent the probability that an examinee is at a certain (hypothetical) state of readiness, having observed his test score and taking the prior expectations into account.

Thus, there is about a 17 percent chance that this crew is combat ready; about a 58 percent chance that they may need some short-term refresher training; and a bit more than a 25 percent chance that they may need extensive retraining. Note that the raw data of 13 hits out of 18 attempts yield a proficiency state of 72 percent. However, when the prior information that only 10 percent of all examinees can be expected to need extensive retraining is used, it becomes more likely that only short-term retraining is needed. Perhaps this is an "almost combat ready" crew that just had a "bad day," rather than being truly non-proficient. Of course, it is ultimately up to the decisionmaker or examiner to interpret these data.

Table 2. Bayesian computation of posterior probabilities for hypothetical states of readiness when test length, test score, and prior probabilities are specified.

1	2	3	4	5
Readiness State, R_x	Prior Probability of Observing R_x	Likelihood* (h= no. hits, n = no. trials)	Prior Times Likelihood	Posterior Probability, or $p(R_x/\text{test score})$
$R_1 = .90$	.60	p(h=13/n=18, R = .90)= .0218	.0131	.167
$R_2 = .80$	.30	p(h=13/n=18, R = .80) = .1507	.0452	.576
$R_3 = .70$	.10	p(h=13/n=18, R = .70) = .2017	.0202	.257
Total	1.00		.0785	1.00

* These values can be found in Hays, W.L. and Winkler, R.L., Statistics (Volume 1), New York: Holt, Rinehart, and Winston, Inc., 1970. Table V, p. 612.

The central characteristic of this model is its ability to use prior information. The more accurate the prior information, the greater the classification accuracy. Further, fewer test trials need be given in order for the examiner to reach a given level of confidence in making classifications. This latter point is very important for weapons crew evaluation, where each live-fire round may cost in the neighborhood of $100. The interaction of test length and quality of prior information is discussed more fully in Novick and Lewis (1974) and Epstein and Steinheiser (1975).

Comparison of Binomial and Bayesian Models

These models are similar in that they take the number of test trials into consideration. Consider a baseball analogy: a pinch hitter may bat only 10 times during a season, yet get three hits. In contrast, some players are able to end the season with an average of about .300, based upon hundreds of times at bat. Clearly, an index of "percent success" must be sensitive to the number of trials!

The binomial model is comparable to classical hypothesis testing procedures and is able to provide estimates for classification errors when the proficiencies, test length, and cutoff scores are specified. Or, it can prescribe a test length to meet the requirements of pre-set acceptable levels of both kinds of classification errors. For example, we could build utilities into the formulation that reflect, among other things, the type and magnitude of classification error for which we are willing to settle. False positive crews may be more acceptable than false negative crews at Fort Hood, but the opposite might be the case in West Germany or Korea, where we might want false positives to be near zero.

The Bayesian model, in contrast, cannot give misclassification probabilities. But it does give the probability of each examinee being at a certain proficiency level. Perhaps more importantly, it is able to use long-term, historical, "prior" information, and can thus exploit the data base which has evolved over time (years, perhaps) about the score distribution of previously tested examinees. To the extent that this prior information is accurate and stable, then classification accuracy will be enhanced, and the number of test trials to reach a specified level of accuracy will be reduced. However, inaccurate prior information can also potentially degrade classification accuracy and may require a larger number of trials to reach a specified level of classification accuracy.

Considering the usefulness of each approach, an appealing alternative would be to use both for test development. Since performance tests are frequently expensive to trial run, particularly for large sample sizes, the use of analytic procedures which have minimal or no data requirements is very desirable. To proceed, the models can be sequenced in the order of their data requirements. Starting with the classical binomial model, the test developer can estimate the required dimensions of the test within the practical constraints of the situation. If good prior information exists, then the Bayesian model can provide more precise information (Novick & Lewis, 1974) and consequent savings in test time and equipment costs.

The application of these models requires attention to the following considerations (Novick & Lewis, 1974). First, the mastery level must be set, either by doctrine or by expert concensus, so that it reflects realistic expectations of performance after training. As the level approaches complete proficiency, the test length must be increased to provide accurate information for classification. Second, the tolerable levels of misclassification must be determined. Of course, the less error tolerance, the more resources necessary for the measurement process and the longer the master test must be. Third, the Bayesian model requires good prior information which can be translated into probability terms. Given stable conditions and quality historical data, the test can be shortened without loss in accuracy. Also, it is more realistic to think of proficiency not as a set of point estimates, but as regions along a continuum. Thus, we may want to say that all examinees who were 90 percent or more proficient should be classified as combat ready; those with proficiencies less than 90 percent but greater than 70 percent should be tested again; and those with proficiencies less than 70 percent should be retrained. These developments are discussed in a paper by Epstein & Steinheiser (1976).

Rasch Model

The model which we shall now consider does not concern itself with minimal numbers of items to make evaluative decisions. Rather, the Rasch

Aligning the metrics of tests which have items in common can be easily accommodated by the model (Kifer, 1976). Figure 1 depicts three Table VIII's given to three different tank units.

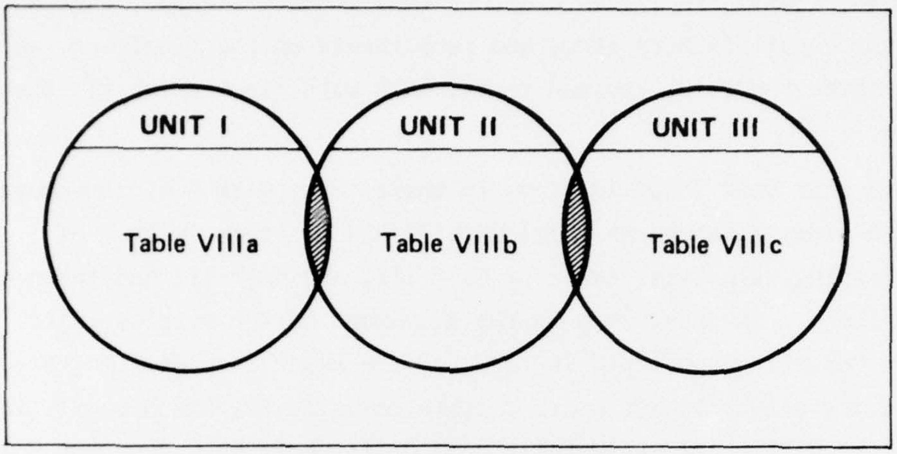


Figure 1. Three Table VIII tests (a, b, and c) with overlapping items given to three tank units (I, II, and III).

As depicted, units I and III are given some items that are in common to Unit II. What we want to be able to do is to compare all three units on the same metric, even though Unit I and Unit III have <u>no items in common</u>.

The items of the Table VIII's would represent some unidimensional aspect of tank gunnery. As indicated in the Boycan and Rose paper (1976), we can conceptualize the tank gunnery domain as eight task families, of which four are main gun tasks. Suppose tank units I, II, and III are tested on proficiency of the tank gunner to fire the main gun in the precision mode (Family V). Using the Rasch model, we would calibrate the items of each of the tests separately. This can be accomplished using a computer program (Mead & Wright, 1976), which is on file at the Army Research Institute computer center. The program produces maximum likelihood estimates of item difficulty and statistics for testing the fit between the Rasch model and the item data.

Item calibration helps us identify problem items. Typically, we examine the reliability and validity indices of trial items in order to determine the final set for measurement. The Rasch model is a form of item analysis, but the emphasis is on the compatibility of the items with the model. Items fail to fit the Rasch scale if the model proves to be too simple to adequately characterize the performance data on the items or if the items measure skills other than the intended skill (Wright & Panchapakesan, 1969). In the case of our tank gunnery example, once we eliminate the unsatisfactory items and recalibrate on the final set, we would have three separate, revised tests, each with items which fit the Rasch model.

Suppose that Unit I had 18 items in their test, with 6 of these overlapping with items taken by Unit II; Unit II had 12 items, with 3 of these overlapping with items taken by Unit III; and Unit III had taken a total of 6 items. The next step in the alignment of the metrics is to estimate an "equating constant" in terms of the Rasch item difficulty indices. Since all three tests are calibrated using the Rasch model, the resulting item difficulty estimates should differ only by a constant. As Kifer (1976) points out, this constant is the difference between the average difficulties for the overlapping items for each group. After identifying the overlapping items that fit the model, we compute the mean difficulties for each of the scales. We then select the metric in which we will express the scores; that is, we could use Unit I as our anchor group for our new scale, or Unit II, or Unit III, or we could define a new metric. Once selected, we calculate the differences among the average difficulties and add the constant to the item difficulties for the groups that require adjustment.

In the tank crew application, we can express the scores on the Unit II metric. If the overlapping six items had the following Rasch item difficulty distributions for units I and II

Item	Unit I	Unit II
a	.53	.43
c	.29	.19
f	.46	.36
h	.41	.31
i	.60	.50
k	.37	.27

then we can easily see that the equating constant in terms of Unit II is -.10 (Avg. Difficulty for Unit I = .44 and the Avg. Difficulty for Unit

II = .34) and that the relationship between the two distributions is perfect (correlation of 1.00). More realistic data are given in Figure 2, where we plot the Rasch item difficulty estimates for Unit I versus Unit II.

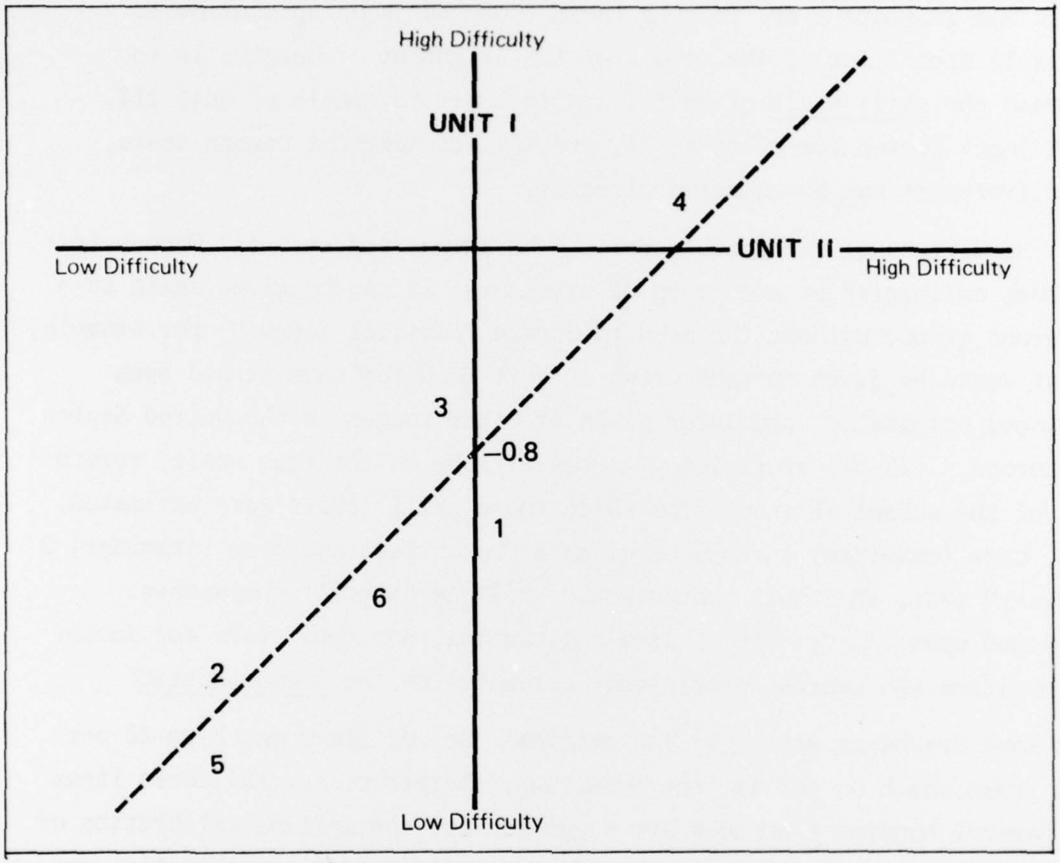


Figure 2. Plot of the Rasch Item Difficulties for Unit I versus Unit II.

If the item selection procedure has worked well, then the relationship will resemble that of the previous example. Although this fit is not so exact as the first, we have depicted a reasonable dispersion. This calls attention to an important consideration in the use of the Rasch model. Although it does not yield estimates of optimal test length, the precision of the model to estimate proficiency depends primarily upon the number of items used in the analysis. As indicated by Wright & Panchapakesan (1969, p. 26), "it is possible to reach any desired level of precision by varying the number of items used in the measurement, just providing that the range of item easiness is reasonably appropriate to the abilities being measured."

The equating constant for Unit I, calculated from the data in Figure 2, is .80. A similar analysis for Unit III might yield a constant of -1.20.* To complete the alignment we add .80 to the Unit I difficulties and -1.20 to the Unit III difficulties. We then re-estimate the skill levels and generate a new scoring table (computer scoring routine is available from Kifer). The result of the alignment of metrics is to increase the skill scale of Unit I and to lower the scale of Unit III. Proficiency scores for units I, II, and III are now on a common scale, and differences can be compared directly.

The virtues of the Rasch model can be summarized as: (1) Once a test has been calibrated on any group of examinees, it can be given again to a different group, without the need to create "parallel forms." For example, a test could be given to tank crews at Fort Hood for whom it had been developed originally, and later given at other ranges in the United States and Europe. (2) All proficiency scores will be on the same scale, regardless of the subset of items from which those skill levels were estimated. Thus, crew (examinee) A could be given a "hard" test and crew (examinee) B an "easy" test, and their scores would still be directly comparable. (3) Based upon the results of item calibration, one could take any subset of the items and express proficiency estimates on the common metric.

Some drawbacks are: (1) The original pool of items may have 20 percent items which do not fit the model and, therefore, several dozen items and several hundred examinees are suggested for the initial calibration of items. (2) Quantitative estimates of misclassification probabilities are not given as output; optimal cutoff scores for a given number of items are not derived.

Conclusions and Applications to Current Testing Programs.

The three approaches are described which ~~outlined above~~ provide different kinds of information relating to the development of a test and use different amounts of information to arrive at their solutions. All relate to the quality of the score that will result. The binomial model yields the probability that

*See Kifer (1976) for an example of the technique with three science achievement tests on three student age groups in Sweden.

examinees will obtain a certain score given a hypothesized "true" level of performance. It provides an initial approximation of test length and cutoff scores without test data. The Bayesian model yields the probability that a particular examinee is a member of a certain proficiency group given a specific score. In this case, prior information on typical performance is combined with the binomial model information to relate the observed score to a true performance level. We therefore require information about the examinee population before the scores are observed. The model can improve the classification accuracy of the test and help us set cutoff points and fix test length at a more efficient level. The Rasch model yields the probability that an examinee in a particular skill group will answer a particular item correctly, given the easiness of the item. We need a good deal of information to accomplish the item calibration and person measurement associated with the Rasch model. Here the emphasis is on the particular items that we will rely on for the measurement; the final set is comprised of items which fit the model.

The Army Research Institute is using these models to analyze test data from two sources; individual pistol marksmanship and tank crew gunnery skills. In the first case, Military Police cadets will fire a .45 caliber pistol at stationary targets, at varying distances and positions. In this rather large experiment, A.R.I. obtained the cooperation of the Military Police School at Fort McClellan, Alabama. Three hundred cadets will fire a total of 240 rounds; this will assure a rich data source for input to the models. In return, A.R.I. will be able to provide specific information to the Military Police School about optimal cutting scores for specific length tests, anticipated misclassification rates, and interpretation of scores once they have been put on a common marksmanship metric. In the second case, the hit-miss scores of 653 tank crews firing main gun rounds through CONUS and USAREUR according to Table VIII are being analyzed. The binomial and Bayesian models are appropriate to the tank gunnery data, because each trial (firing a main gun round) is quite costly. The task of the decisionmaker is to make evaluative (i.e., "pass-fail," or "combat ready" vs. "not combat ready") judgments on the basis of test score data, using as few trials as possible. Subject to these constraints, the binomial and Bayesian models will tell the decisionmaker what his chances are of making correct and incorrect classifications. The logistic model discussed in the last section of this paper will also make a contribution to the evaluative process, but in a different manner than the first two models. This model would allow scores to be placed on the same "gunnery" metric, even though the tests were conducted using different ranges.

References

Boycan, G. & Rose, A. An analytic approach to estimating the generalizability of crew performance objectives. Paper presented at the annual meeting of the Military Testing Association, Gulf Shores, Alabama, Oct. 1976.

Epstein, K.I., & Steinheiser, F.H. Comparison of error rates and misclassification probabilities using binomial and Bayesian models for personnel classification. Paper presented at the 22nd Conference on the Design of Experiments in the Army Research, Development and Testing, held at Harry Diamond Laboratories, 20-22 Oct. 1976.

________________. A Bayesian method for evaluating trainee proficiency. In 8th Naval Training Equipment Center/Industry Conference Proceedings. NTEC: Orlando, Florida, 19-20 Nov. 1975.

Hays, W.L. & Winkler, R.L. Statistics (Volume 1). New York: Holt, Rinehart, & Winston, 1970.

Kifer, E. Estimating scores on a common metric using the Rasch model: An I.E.A. application. Paper presented at the annual A.E.R.A. meeting, San Francisco, April 1976.

Macready, G.B., Epstein, K.I., Steinheiser, F.H., & Mirabella, A. Methods and models for criterion-referenced testing. Army Research Institute, Technical Paper, in press, 1976.

Mead, R., & Wright, B. Using the Rasch model for analyzing behavioral data. Interim Report to the Army Research Institute from the University of Chicago, Nov. 1976.

Millman, J. Passing scores and test lengths for domain-referenced tests. Review of Educational Research, 1973, 43, 205-215.

Novick, M., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Center for the Study of Evaluation Monograph Series in Evaluation, III: Problems in criterion-referenced measurement. Los Angeles, Calif: Center for the Study of Evaluation at U.C.L.A., 1974.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche, 1960 (English translation).

________. On general laws and the meaning of measurement in psychology. In Proceedings of the fourth Berkeley Symposium on mathematical statistics. Berkeley: University of California Press, 1961, IV, 321-334.

________. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Whitely, S.E., & Dawis, R.V. The nature of objectivity with the Rasch model. *Journal of Educational Measurement*, 1974, *11*, 163-168.

Wright, B.D., & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, *29*, 23-48.